DESIGN & IMPLEMENTATION OF DATA WAREHOUSE PROTOTYPE WITH IN THE CONTEXT OF RELATIONAL ONLINE ANALYTICAL PROCESSING (DATA ANALYSIS)

Dr. Vivek Chaplot,

B.N. PG College, Udaipur.

1 INTRODUCTION

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject. A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product. Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer. Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered. A data warehouse is a copy of transaction data specifically structured for query and analysis. This is a functional view of a data warehouse. Kimball did not address how the data warehouse is built like Inmon did, rather he focused on the functionality of a data warehouse. After the tools and team personnel selections are made, the data warehouse design can begin. The following are the typical steps involved in the data warehousing project cycle.

- Requirement Gathering
- Physical Environment Setup
- Data Modeling
- ETL
- OLAP Cube Design
- Front End Development
- Report Development
- Performance Tuning
- Query Optimization
- Quality Assurance
- Rolling out to Production
- Production Maintenance
- Incremental Enhancements

Each page listed above represents a typical data warehouse design phase, and has several sections:

- **Task Description:** This section describes what typically needs to be accomplished during this particular data warehouse design phase.

- **Time Requirement:** A rough estimate of the amount of time this particular data warehouse task takes.

- **Deliverables:** Typically at the end of each data warehouse task, one or more documents are produced that fully describe the steps and results of that particular task. This is especially important for consultants to communicate their results to the clients.

- **Possible Pitfalls:** Things to watch out for. Some of them obvious, some of them not so obvious. All of them are real.

Requirement Gathering

The first thing that the project team should engage in is gathering requirements from end users. Because end users are typically not familiar with the data warehousing process or concept, the help of the business sponsor is essential. Requirement gathering can happen as one-to-one meetings or as Joint Application Development (JAD) sessions, where multiple people are talking about the project scope in the same meeting.

The primary goal of this phase is to identify what constitutes as a success for this particular phase of the data warehouse project. In particular, end user reporting / analysis requirements are identified, and the project team will spend the remaining period of time trying to satisfy these requirements.

Associated with the identification of user requirements is a more concrete definition of other details such as hardware sizing information, training requirements, data source identification, and most importantly, a concrete project plan indicating the finishing date of the data warehousing project.

Based on the information gathered above, a disaster recovery plan needs to be developed so that the data warehousing system can recover from accidents that disable the system. Without an effective backup and restore strategy, the system will only last until the first major disaster, and, as many data warehousing DBA's will attest, this can happen very quickly after the project goes live.

Deliverables

- A list of reports / cubes to be delivered to the end users by the end of this current phase.

- A updated project plan that clearly identifies resource loads and milestone delivery dates.

Possible Pitfalls

This phase often turns out to be the trickiest phase of the data warehousing implementation. The reason is that because data warehousing by definition includes data from multiple sources spanning many different departments within the enterprise, there are often political battles that center on the willingness of information sharing. Even though a successful data warehouse benefits the enterprise, there are occasions where departments may not feel the same way. As a result of unwillingness of certain groups to release data or to participate in the data warehousing requirement definition, the data warehouse effort either never gets off the ground, or could not start in the direction originally defined.

When this happens, it would be ideal to have a strong business sponsor. If the sponsor is at the CXO level, she can often exert enough influence to make sure everyone cooperates.

Deliverables

- A list of reports / cubes to be delivered to the end users by the end of this current phase.

- A updated project plan that clearly identifies resource loads and milestone delivery dates.

Possible Pitfalls

This phase often turns out to be the trickiest phase of the data warehousing implementation. The reason is that because data warehousing by definition includes data from multiple sources spanning many different departments within the enterprise, there are often political battles that center on the willingness of information sharing. Even though a successful data warehouse benefits the enterprise, there are occasions where departments may not feel the same way. As a result of unwillingness of certain groups to release data or to participate in the data warehousing requirement definition, the data warehouse effort either never gets off the ground, or could not start in the direction originally defined.

When this happens, it would be ideal to have a strong business sponsor. If the sponsor is at the CXO level, she can often exert enough influence to make sure everyone cooperates.

Physical Environment Setup

Once the requirements are somewhat clear, it is necessary to set up the physical servers and databases. At a minimum, it is necessary to set up a development environment and a production environment. There are also many data warehousing projects where there are three environments: Development, Testing, and Production.

It is not enough to simply have different physical environments set up. The different processes (such as ETL, OLAP Cube, and reporting) also need to be set up properly for each environment.

It is best for the different environments to use distinct application and database servers. In other words, the development environment will have its own application server and database servers, and the production environment will have its own set of application and database servers.

Having different environments is very important for the following reasons:

- All changes can be tested and QA'd first without affecting the production environment.
- Development and QA can occur during the time users are accessing the data warehouse.
- When there is any question about the data, having separate environment(s) will allow the data warehousing team to examine the data without impacting the production environment.

Deliverables

- Hardware / Software setup document for all of the environments, including hardware specifications, and scripts / settings for the software.

Report Development

- Report specification typically comes directly from the requirements phase. To the end user, the only direct touchpoint he or she has with the data warehousing system is the reports they see. So, report development, although not as time consuming as some of the other steps such as ETL and data modeling, nevertheless plays a very important role in determining the success of the data warehousing project.
- One would think that report development is an easy task. How hard can it be to just follow instructions to build the report? Unfortunately, this is not true. There are several points the data warehousing team need to pay attention to before releasing the report.
- **User customization:** Do users need to be able to select their own metrics? And how do users need to be able to filter the information? The report development process needs to take those factors into consideration so that users can get the information they need in the shortest amount of time possible.
- **Report delivery:** What report delivery methods are needed? In addition to delivering the report to the web front end, other possibilities include delivery via email, via text messaging, or in some form of spreadsheet. There are reporting solutions in the marketplace that support report delivery as a flash file. Such flash file essentially acts as a mini-cube, and would allow end users to slice and dice the data on the report without having to pull data from an external source.

CONCLUSIONS

Successful data management is an important factor in developing support systems for the decision-making process. Traditional database systems, called operational or transactional, do not satisfy the requirements for data analysis of the decision-making users. An operational database supports daily business operations and the primary concern of such database is to ensure concurrent access and recovery techniques that guarantee data consistency. Operational databases contain detailed data, do not include historical data, and since they are usually highly normalized, they perform poorly for complex queries that need to join many relational tables or to aggregate large volumes of data.

A DW represents a large repository of integrated and historical data needed to support the decision-making process. The structure of a DW is based on a multidimensional model. This model includes measures that are important for analysis, dimensions allowing the decision-making users to see these measures from different perspectives, and hierarchies supporting the presentation of detailed or summarized measures. The characteristics of a multidimensional model specified for the DW can be applied for a smaller structure, a data mart, which is different from the DW in the scope of its analysis. A data mart refers to a part of an organization and contains limited amount of data.

DWs have become the main technology for DSS. DSSs require not only the data repository represented by DW, but also the tools that allow analysing data. These tools include different kinds of applications; for example, software that include statistics and data mining techniques offers complex analysis for a large volume of data to identify profiles, behaviour, and tendencies. On the other hand, OLAP tools can manage high volumes of historical data allowing for dynamic data manipulations and flexible interactions with the end-users through the drill-down, roll-up, pivoting, and slicingdicing operations. Furthermore, OLAP tools are based on multidimensional concepts similar to DW multidimensional model using for it measures, dimensions, and hierarchies. If the DW data structure has a well-defined multidimensional model, it is easier to fully exploit OLAP tools capabilities.

In this thesis, widely accepted conceptual and logical design approaches in DW design are discussed. In the conceptual design phase DF, starER, ME/R and OOMD design models are compared. OO design model is significantly better than the other design approaches. OOMD supports conceptual design phase with a rich set of diagrams that enables the designer model all the business information and requirements using a case tool with UML. OOMD design model meets the following factors while the others lack one or more:

- Additivity of measures
- Many-to-many relationships with dimensions
- Derived measures
- Nonstrict and complete classification hierarchies
- Categorization of dimensions (specialization/generalization)
- Graphic notation
- Specifying user requirements
- Case tool support

In the logical design phase flat, terraced, star, fact constellation, galaxy, snowflake, star cluster and star flake schemas are discussed. Among these logical design models, star schema, snowflake schema and the fact constellation schema are the mostly used models commercially. These three models are compared in terms of efficiency, usability, reusability and flexibility quality factors among which efficiency the most important one is considering DW modeling. Considering these factors and the requirements of the business and considering the trade-off between redundancy and the query performance, either snowflake or star schema may be the best choice in the design.

REFERENCES

1. Romm M., Introduction to Data Warehousing, San Diego SQL User Group
2. Goyal N., Introduction to Data Warehousing, BITS, Pilani Lecture Notes
3. Franconi E., Introduction to Data Warehousing, Lecture Notes, http://www.inf.unibz.it/~franconi/teaching/2002/cs636/2 ,2002
4. Pang L., Data Warehousing and Data Mining, Leslie Pang Web Site and Lecturer Notes
5. Gatziu S. and Vavouras A., Data Warehousing: Concepts and Mechanisms, 1999
6. Thomas Connolly & Carolyn Begg., "Database Systems, 3th Edition", Addison- Wesley, 2002
7. Gatierrez A. and Marotta A., An Overview of Data Warehouse Design Approaches and Techniques, Uruguay, 2000
8. Reed Jacobson., "Microsoft® SQL Server 2000 Analysis Services", ISBN 0-7356-0904-7, 2000
9. Rizzi S., Open Problems in Data Warehousing., http://sunsite.informatik.rwthaachen.de/Publications/CEUR-WS/Vol-77/ DMDW 2003, Berlin, Germany
10. J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Chapter2: Data Warehouse and OLAP Technology for Data Mining, Barnes & Nobles, 2000
11. W. H. Inmon, "Building the Data Warehouse, 3th Edition", John Wiley, 2002
12. Moody D. L. and Kortink M. A. R., From Enterprise Models to Dimensional Models: Methodology for Data Warehouse and Data Mart Design, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-28/DMDW 2000 , Stockholm, Sweden
13. Tryfona N., Busborg F., Christiansen J. G., starER: A Conceptual Model for Data Warehouse Design, Proceeding of the ACM 2nd International Workshop Data Warehousing and OLAP (DOLAP99), 1999
14. Sapia C., Blaschka M., Höfling G., Dinter B., Extending the E/R Model for the Multidimensional Paradigm, Proceeding 1st International Workshop on Data Warehousing and Data Mining (DWDM98), 1998
15. Golfarelli M., Maio D., Rizzi S., Conceptual Design of Data Warehouses from E/R Schemas, Proceeding of the 31st Hawaii International Conference on System Sciences (HICSS-31), Vol. VII,1998
16. Golfarelli M., Maio D., Rizzi S., The Dimensional Fact Model: A Conceptual Model For Data Warehouses, International Journal of Cooperative Information Systems (IJCIS), Vol. 7, 1998
17. Golfarelli M, Rizzi S., A Methodological Framework for Data Warehouse Design, Proceeding of the ACM DOLAP98 Workshop, 1998
18. Lujan-Mora S., Trujillo J., Song I., Multidimensional Modeling with UML Package Diagrams, 21st International Conference on Conceptual Modeling (Er2002), 2002
19. Trujillo J., Palomar M., An Object Oriented Approach to Multidimensional Database Conceptual Modeling (OOMD) , Proceeding 1st International Workshop on Data Warehousing and OLAP (DOLAP98), 1998
20. Kimball R., http://www.dbmsmag.com/9708d15.html "A Dimensional Modeling Manifesto", DBMS Magazine, Aug 1997
21. Kimball R., "The Data Warehouse Toolkit", John Wiley, 1996
22. Martyn T., Reconsidering Multi-Dimensional Schemas, SIGMOD Record, Vol.